Lead Scoring Case Study

Predicting Potential Leads Using Logistic Regression

Presented by:

Gagan D

Problem Statement

OBJECTIVE:

Identify potential leads who are most likely to convert into paying customers.

BUSINESS NEED:

- X education is facing the challenge of efficiently allocating marketing resources.
- Prioritizing high-potential leads will enable more effective customer acquisition strategies.

CHALLENGES:

- Large volume of leads makes it difficult to manually identify high-quality ones.
- Without a systematic scoring method, marketing efforts could be wasted on low-probability leads.

Data Preprocessing

Handling Missing Data:

- **Imputation:** Missing values were imputed using [mean, median, mode, or removed rows, depending on your process].
- **Removed Records:** Rows with significant missing data were removed to maintain dataset integrity.

Data Transformation:

- **Encoding Categorical Variables:** Categorical features such as Lead Source, Country, etc., were converted into numerical form using dummy variables (one-hot encoding).
- **Scaling Features:** Numerical features like Total Time Spent on the Website were scaled using standardization or normalization to ensure that different scales don’t bias the model.

Feature Selection:

- **Important Features Identified:** Lead Source, Time Spent on Website, Page Views, etc., were found to be the most impactful features for predicting lead conversion.
- **Dimensionality Reduction:** Any dimensionality reduction technique like PCA was used to reduce the number of features.

Model Selection and Building

Model Used:

- **Logistic Regression** was selected as the model for binary classification (converted vs. not converted).

Why Logistic Regression

- **Simplicity and Interpretability:** Logistic regression is easy to interpret and understand, making it suitable for business stakeholders.
- **Handling Binary Outcomes:** It directly predicts probabilities for the two possible outcomes—conversion or no conversion.

Technical Details:

- **Train-Test Split:** The dataset was split into training and test sets using an [80-20 or 70-30] ratio.
- **Regularization:** L1 or L2 regularization was applied to avoid overfitting.
- **Cross-Validation:** K-fold cross-validation was used to validate the model performance on different subsets of the data.

Model Evaluation

Performance Metrics:

Accuracy: The percentage of correctly predicted outcomes.

Precision: The proportion of true positive predictions among all positive predictions.

Recall: The proportion of actual positive cases that were predicted correctly.

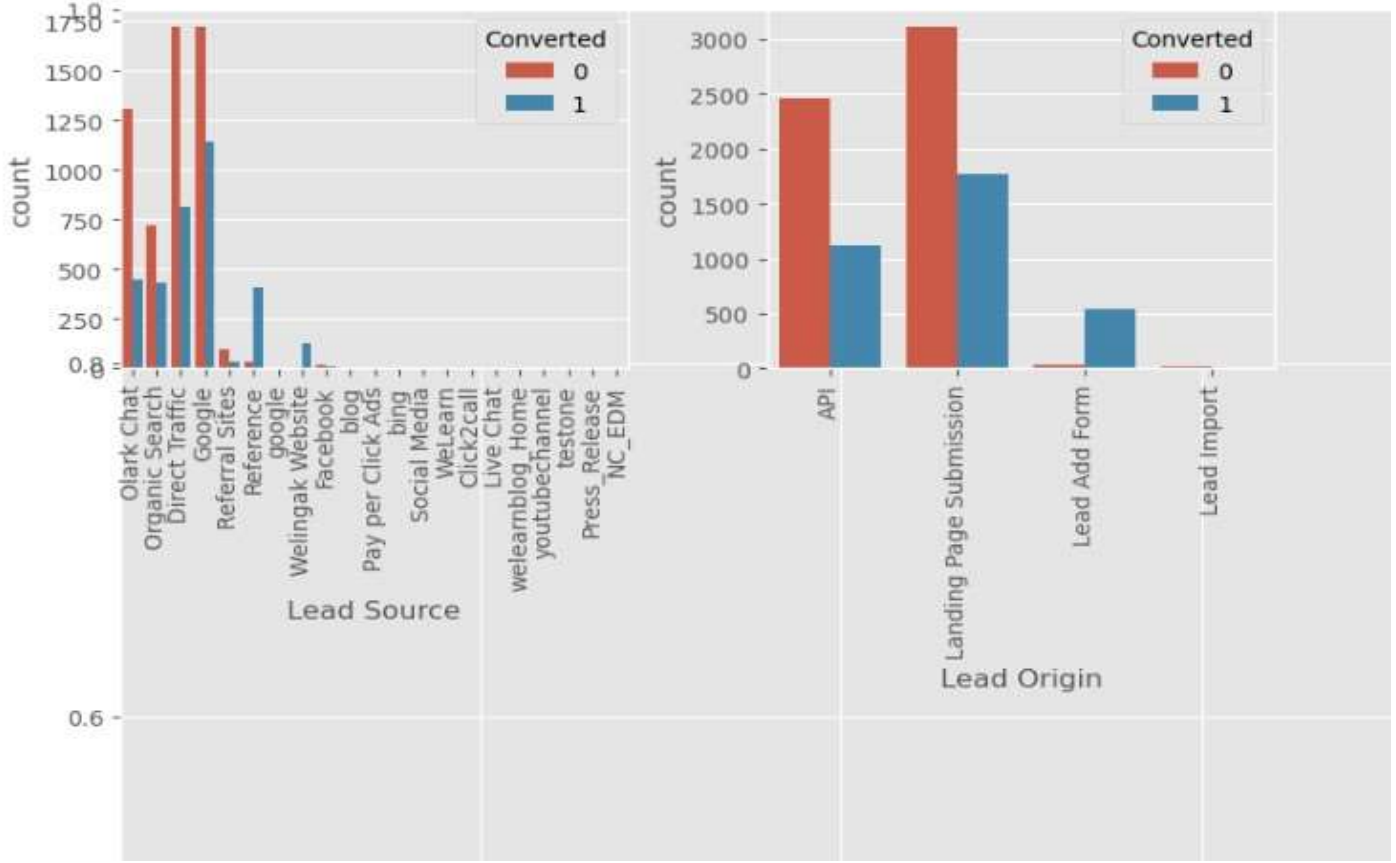
F1-Score: The harmonic mean of precision and recall.

ROC-AUC Score: A measure of how well the model distinguishes between classes (1 = perfect, 0.5 = random).

Confusion Matrix:

A confusion matrix was generated to provide insights into the model's true positive, true negative, false positive, and false negative predictions.

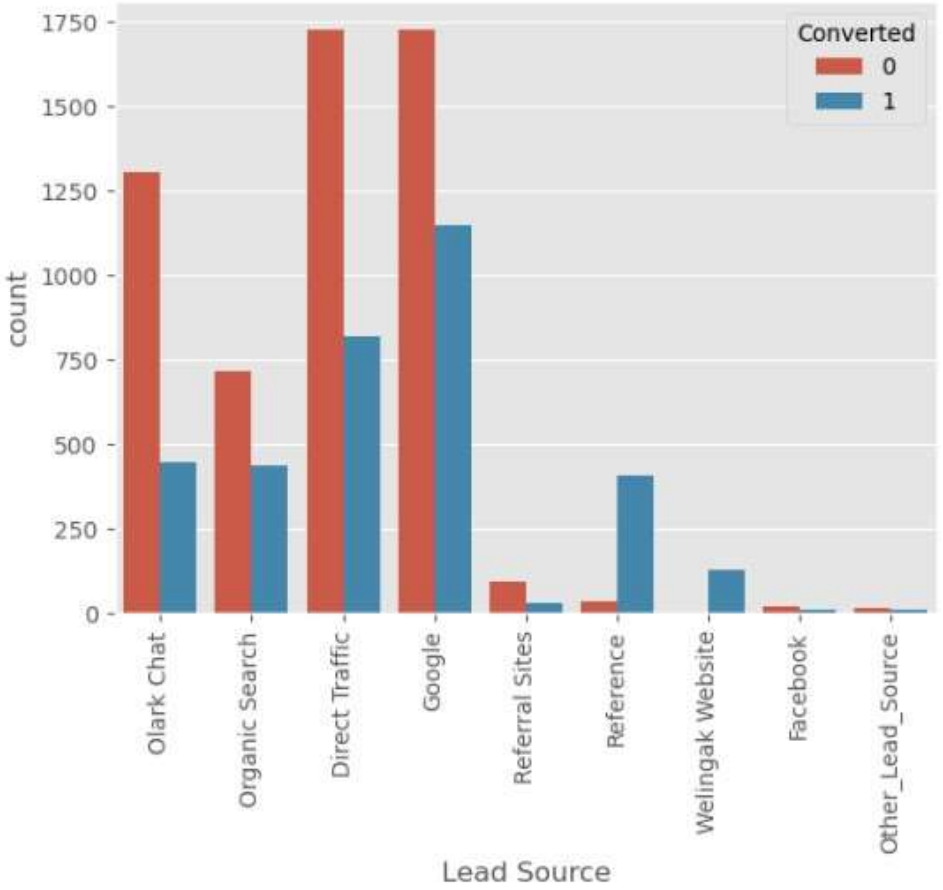
Customer Conversion by Last Activity



This bar chart illustrates how different last activities affect the likelihood of conversion. The two colors represent converted (blue) and non-converted (red) customers. Key insights include:

- Activities like "Email Opened" and "SMS Sent" show high conversion rates.
- Certain actions like "Page Visited on Website" have higher non-conversion counts. This analysis helps prioritize effective communication channels.

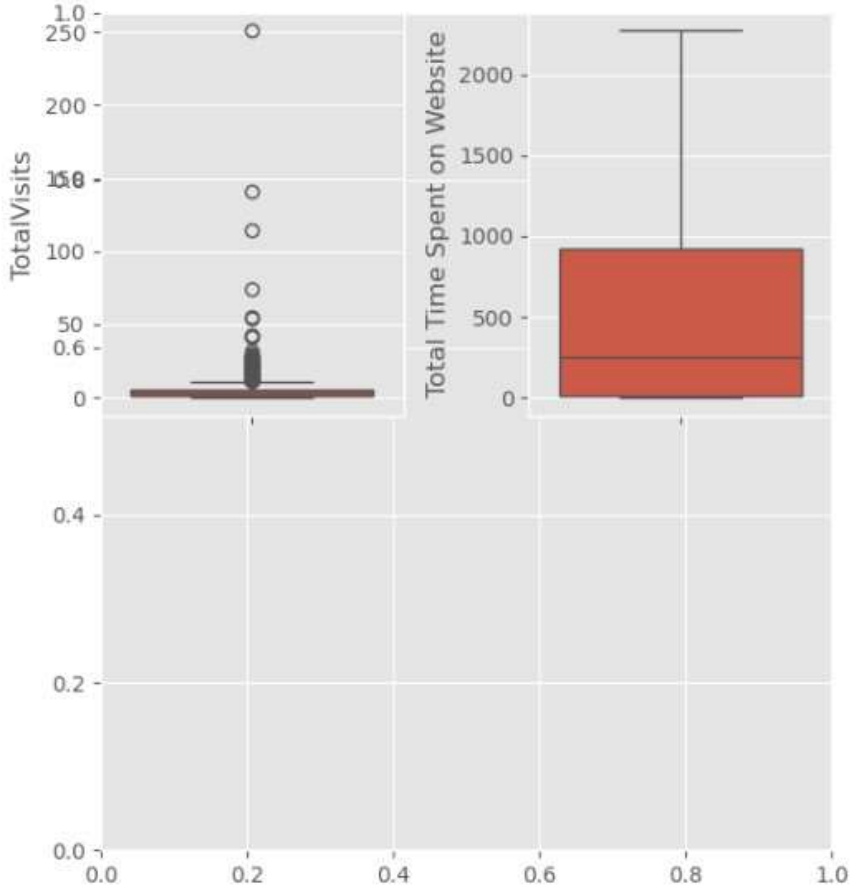
Impact of Total Visits and Time Spent on Conversion



The box plots depict the distribution of total website visits and time spent on the website for both converted and non-converted customers. Key findings include:

- Customers who converted tend to have more website visits and spend significantly more time on the website.
- Higher total time spent on the website strongly correlates with increased conversion probability. This indicates that engagement with the website is a crucial factor in conversion.

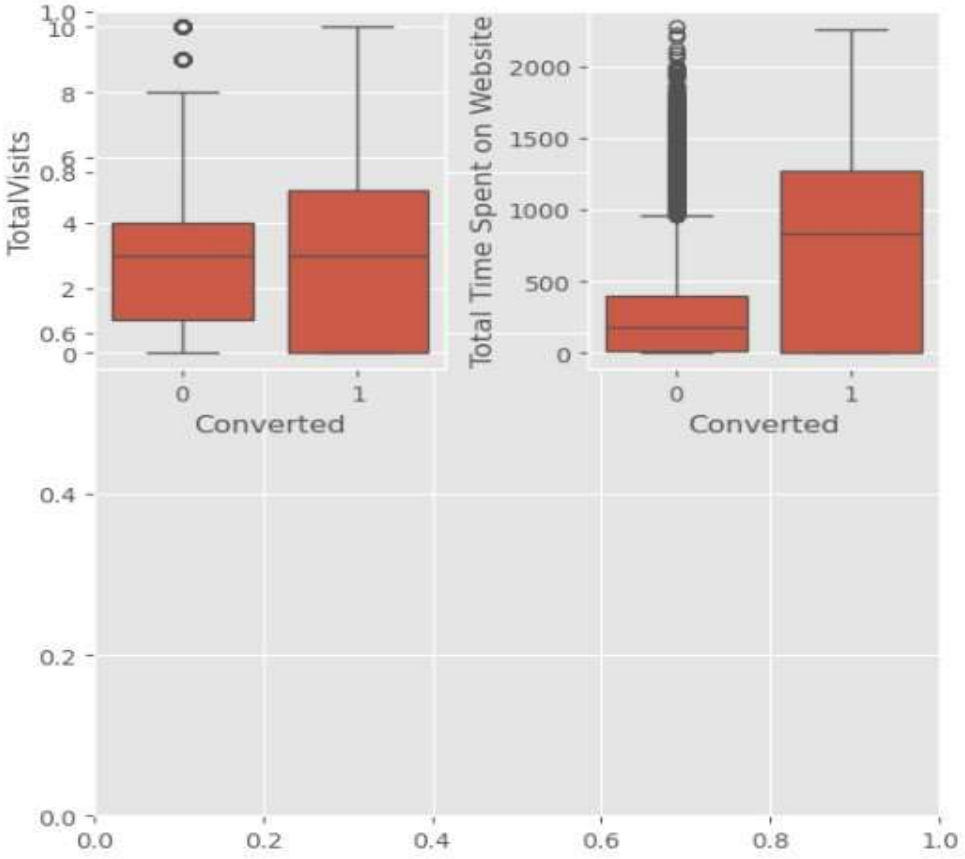
Conversion Rates by Lead Source



This bar chart compares the conversion rates across various lead sources. The red bars represent non-converted leads, while the blue bars represent converted leads. Key insights:

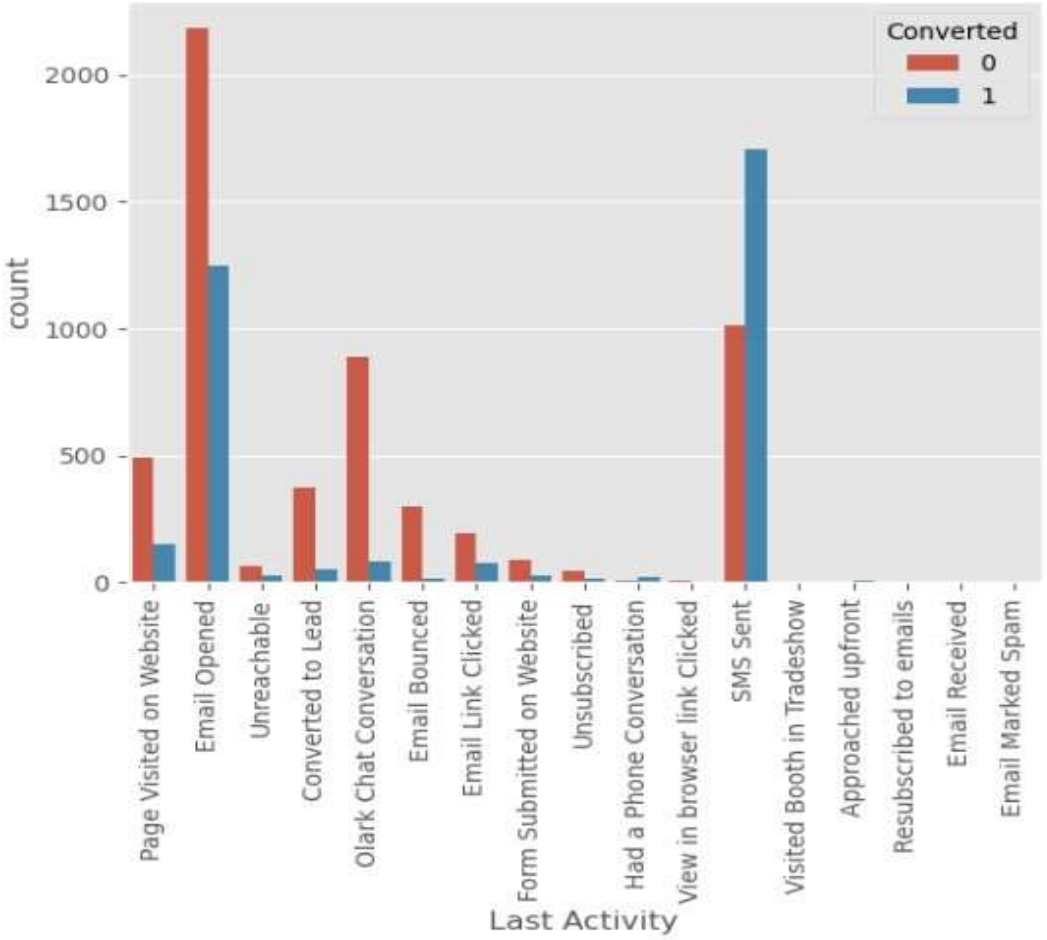
- Direct traffic and Google are major lead sources, with Direct Traffic showing higher conversions.
- Organic search and referral sites contribute to a smaller portion of conversions. This highlights the most effective marketing channels for driving conversions.

Lead Conversion Analysis: Impact of Total Visits and Time Spent on Website



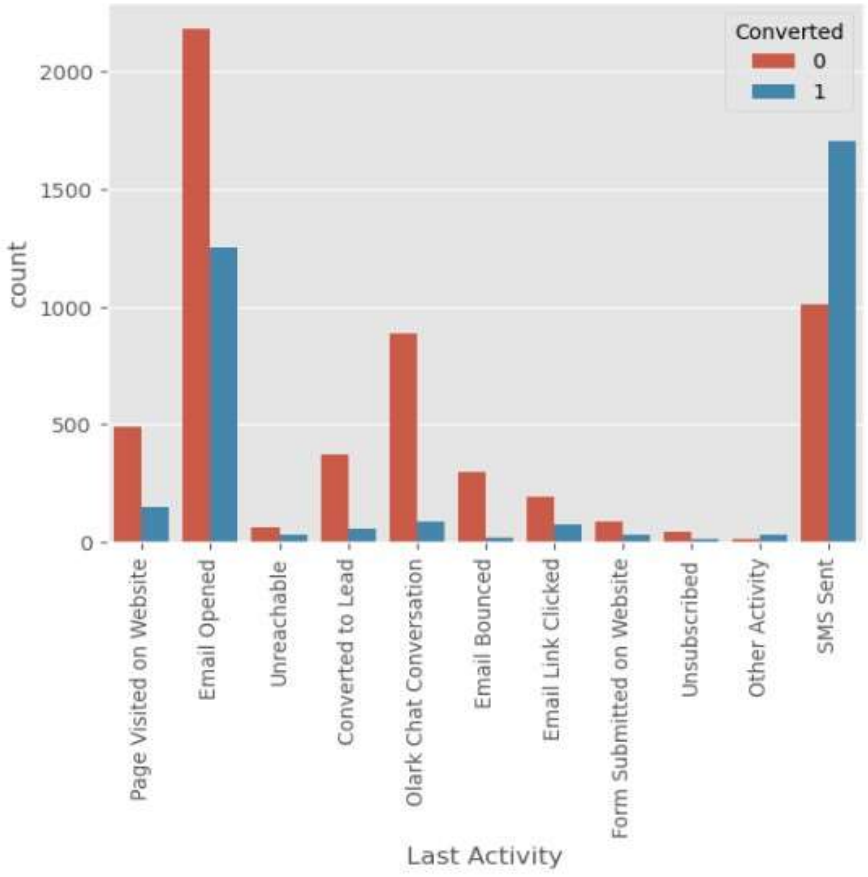
This chart compares the total number of visits and the total time spent on the website between leads that converted (1) and those that did not (0). The box plot for total visits shows that converted leads tend to visit more frequently, with the median value higher than non-converted leads. Similarly, the box plot for total time spent shows that converted leads spend significantly more time on the website, with a notable difference in the interquartile range. These insights suggest that both visit frequency and engagement time are important indicators of conversion likelihood.

Lead Activity Breakdown: Conversion Trends Across Last Activities



This bar chart shows the distribution of last activities performed by leads before conversion or non-conversion. The most common activities include "Email Opened," which shows a large number of non-converted leads, and "SMS Sent," where converted leads are dominant. The chart also highlights activities such as "Form Submitted on Website" and "Olark Chat Conversation," which appear to contribute more positively to conversions. These insights can help identify key touchpoints that drive lead conversion, allowing for optimization of marketing efforts around these activities.

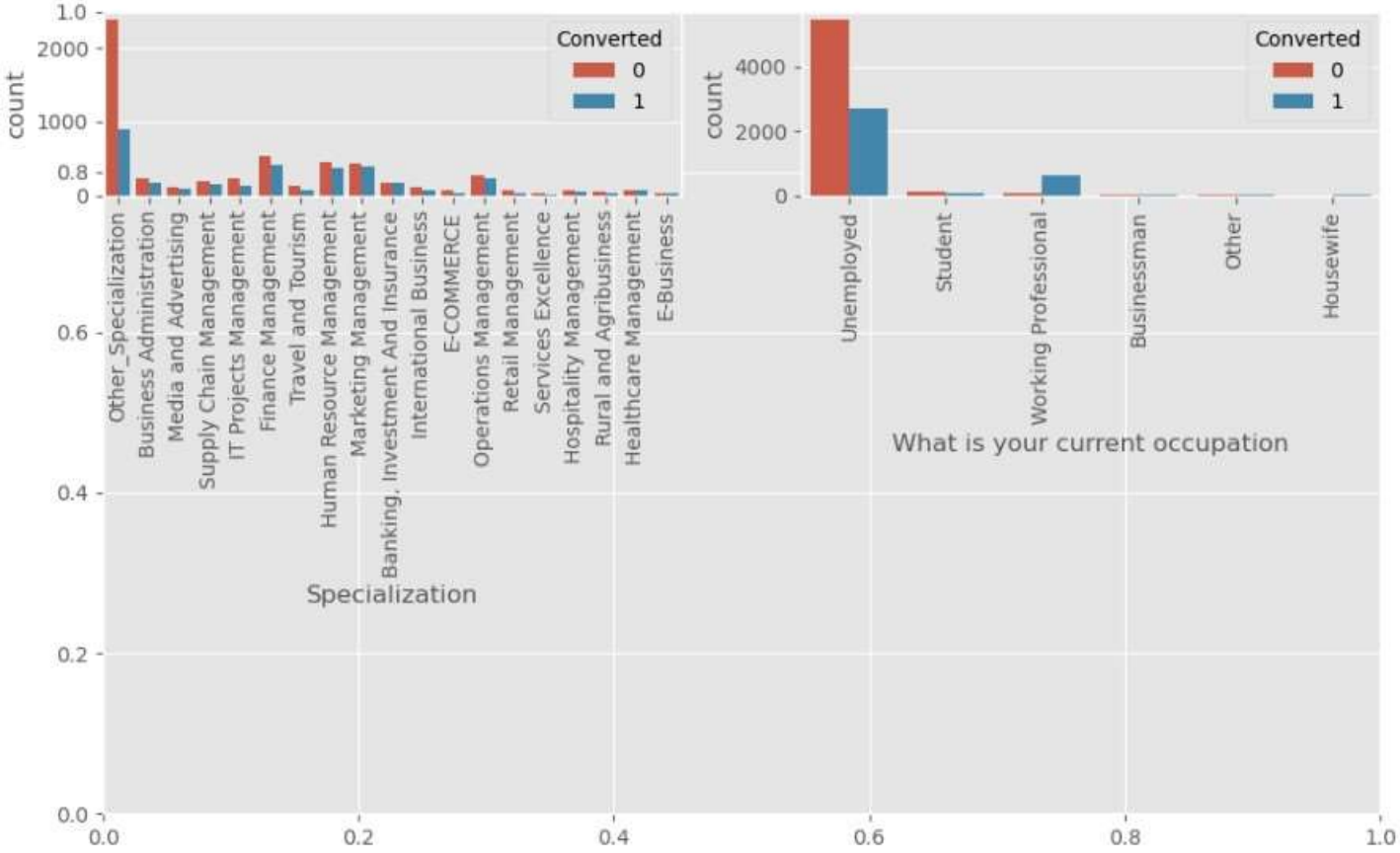
Lead Quality and Tag Distribution for Conversion Rates



This graph presents the relationship between **lead quality** and **tags** (reasons for lead engagement) with respect to the lead conversion outcome.

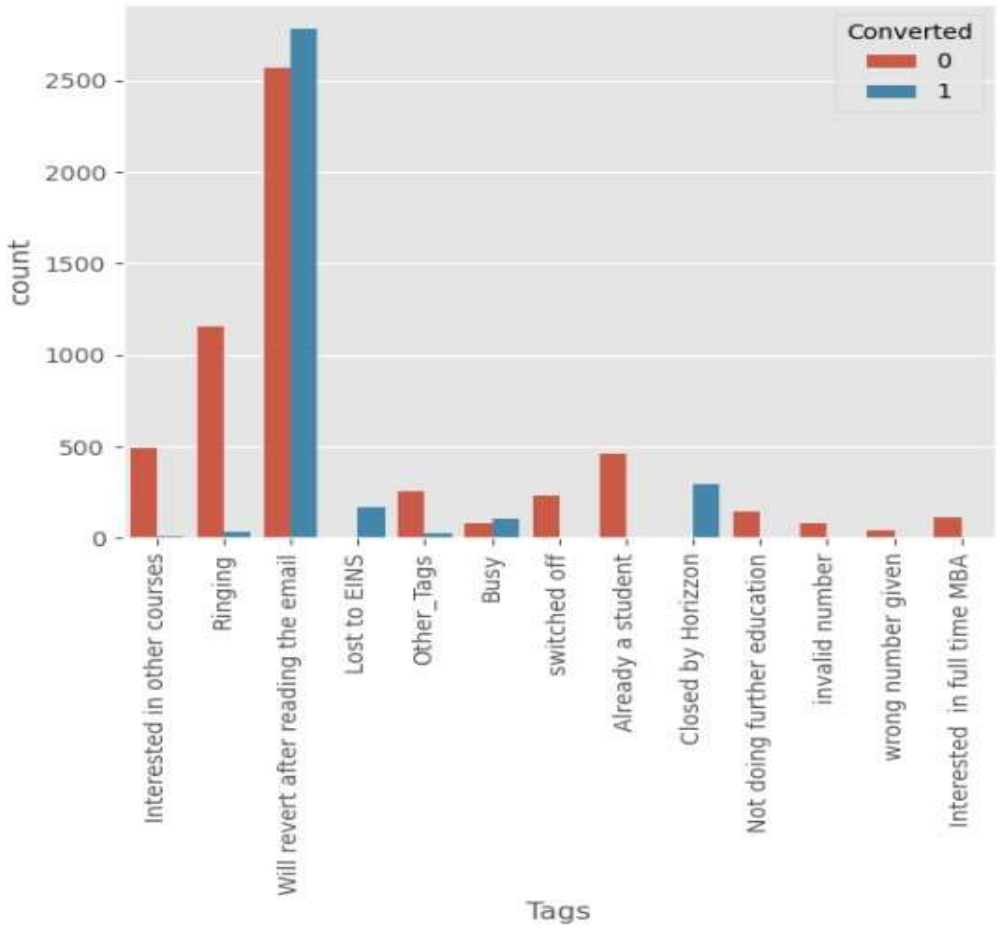
- The left side represents different **lead quality categories** such as "Low in Relevance," "Not Sure," "High in Relevance," etc., with counts of converted (blue) and non-converted (red) leads.
- A significant portion of the "Not Sure" category has not converted, highlighting the need for further qualification.
- The right side provides a detailed look at **tags** (reasons given by the leads for engagement or disengagement), showing the distribution of converted (blue) and non-converted (red) leads.
- The majority of leads tagged as "Interested in other courses" or "Will never after reading the email" are not converting, indicating areas where the business should reconsider their targeting strategy.
- This analysis helps to identify the key factors that influence conversion rates and guide marketing and outreach strategies to improve lead quality and engagement.

Analysis of Lead Conversion Based on Specialization and Occupation



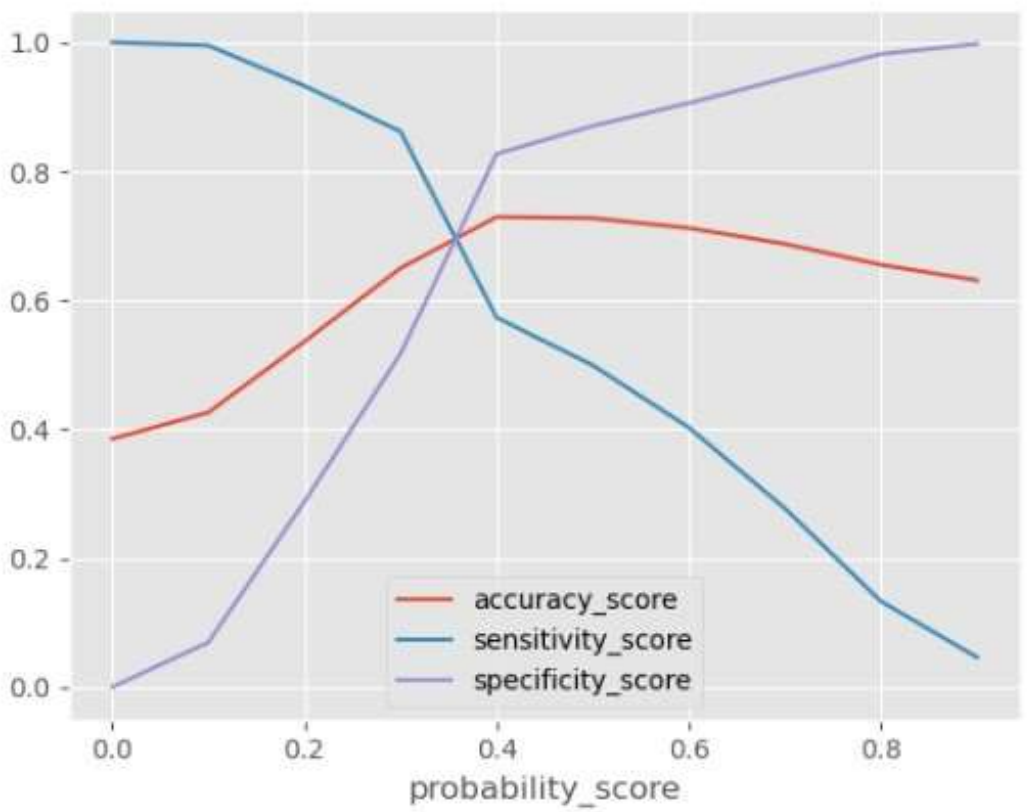
This graph provides a comparative analysis of lead conversion rates categorized by specialization and current occupation. The left section represents how different specializations, such as Business Administration, Supply Chain Management, and Media, influence lead conversion. The right section focuses on the influence of occupation types like "Unemployed," "Student," or "Working Professional." The color-coding shows converted (1) and non-converted (0) leads, highlighting trends where higher conversion occurs, especially among unemployed individuals with diverse specializations. This insight aids in identifying key demographics for targeting effective lead conversion strategies.

Impact of Lead Tags on Conversion Rates



This graph explores the influence of various tags on lead conversion outcomes. Tags such as "Will revert after reading the email," "Interested in other courses," and "Ringing" have been categorized by conversion status (0: Not Converted, 1: Converted). Notably, the tag "Will revert after reading the email" shows a significant conversion rate, while other tags like "Lost to EINS" and "Ringing" have higher non-conversion counts. This analysis provides a clear understanding of how specific actions or conditions represented by tags affect the likelihood of lead conversion, enabling better targeted follow-ups and improved conversion strategies.

ROC Curve for Model Performance at Varying Probability Thresholds



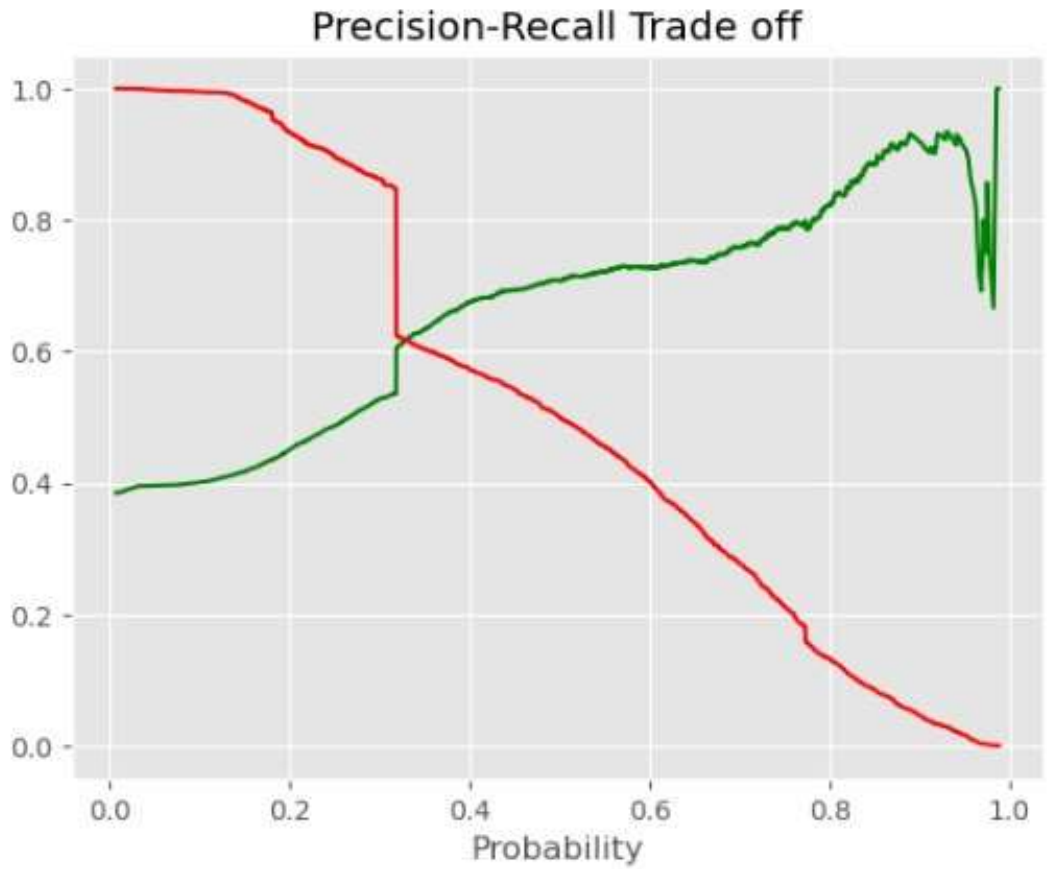
This graph shows the performance of the model across various probability thresholds for lead conversion prediction. The **accuracy score** (red), **sensitivity score** (blue), and **specificity score** (purple) change as the decision threshold varies from 0 to 1.

•**Accuracy** remains relatively stable around the middle but begins to decline as the threshold moves towards the extremes.

•**Sensitivity** (true positive rate) decreases significantly as the threshold increases, indicating fewer true leads are correctly identified.

•**Specificity** (true negative rate), in contrast, improves as the threshold increases, highlighting the model's increasing ability to reject false leads.

Precision-Recall Trade-off Curve



This graph visualizes the precision-recall trade-off across different probability thresholds for a classification model. The green line represents recall, while the red line represents precision. As the probability threshold increases, precision rises and recall decreases, demonstrating the inverse relationship between the two metrics. The curve helps in determining the optimal threshold where a balance between precision and recall can be achieved, depending on the model's application. This analysis is critical in fine-tuning model performance for tasks where either precision or recall is prioritized, such as lead conversion predictions.

Key Findings

Top Features Affecting Lead Conversion:

- **Lead Source:** Leads from specific sources (e.g., direct traffic, referral) were more likely to convert.
- **Total Time Spent on Website:** Leads who spent more time on the website had a higher probability of conversion.
- **Number of Page Views:** Higher page views were correlated with increased conversion likelihood.
- **Activity Engagement:** Leads who interacted with webinars, emails, or downloads had higher chances of converting.

Business Insights:

- **High-Quality Lead Sources:** Marketing efforts should focus on lead sources that show high conversion rates.
- **Website Engagement:** Enhancing engagement (e.g., increasing time spent on the website) can improve conversion rates.
- **Personalized Outreach:** Targeted marketing strategies based on customer activity (like webinars or email clicks) can enhance conversions.

Business Recommendations

Optimizing Lead Scoring Process:

- Focus marketing efforts on high-scoring leads that are more likely to convert.
- Automate lead prioritization using the logistic regression model to ensure sales teams focus on the most promising leads.

Targeted Marketing Strategies:

- **Personalized Marketing:** Customize marketing messages based on lead behavior and engagement (e.g., more time spent on the website or specific lead sources).
- **Resource Allocation:** Direct resources towards lead sources that are more likely to convert, maximizing ROI on marketing spend.

Potential Business Impact:

- **Increased Conversion Rates:** By focusing on high-potential leads, the company can increase overall sales conversions.
- **Improved Customer Satisfaction:** Engaging leads with personalized outreach can improve the customer experience and long-term brand loyalty.
- **Cost Efficiency:** Reducing time and resources spent on low-potential leads will result in more efficient marketing campaigns.

Conclusion

Summary of Findings:

- A logistic regression model was developed to predict lead conversion, helping prioritize marketing efforts.
- Key factors influencing conversion include lead source, time spent on the website, and user activity engagement.

Business Impact:

- Implementing the model can lead to increased efficiency in lead management.
- The business can expect improved conversion rates by targeting high-potential leads.

Future Scope:

- **Model Enhancement:** Further refinement of the model by incorporating more features or testing advanced machine learning algorithms.
- **Continuous Monitoring:** Regular updates and monitoring of the model performance to ensure it stays accurate over time.

Thank You